State of Connecticut

Data Management and Data Warehouse Domain Technical Architecture

June 6, 2002

Version 2.0

History of Changes

06/06/2002	Revised definitions for component technologies: Multi-Dimensional Databases (MDDBMS; Data Warehouses – Data Marts; Operational Data Stores Reorganized components into two major Data Management and Data Warehouse sections Addition to Standards section to reflect ETL tool under research (new table 7) Modification to Best Practices Data (new 5), Best Practices Databases (deleted 4), Best Practices Data Warehouse (deleted old 1, 4, 5, 14, 15, 19, 22 and renumbered remaining, revised 2, revised 3 and 4 rationale, revised 7 and 8) , Best Practices Data Replication (new 1, deleted old 2, revised new 2), Best Practices Metadata Repository (Combined old 1 and 2, revised new 2) Modifications to Principles 1 (clarification of terminology in implications), 5 (implications), 8 (implications),13 (implications), 15 (rationale) Addition of Principles 7,11 with renumbering of existing principles Format and grammar modifications
01/03/2001	Format modifications
12/06/2000	"To Be Determined" section added
11/22/2000	In Table 3Desktop (PC) Database reassigned Oracle 8 and IBM DB2 (UDB) to research category. In Table 5 Database Connectivity Standards added ADO, DAO and RDO to standards list;

Table of Contents

Mission

The Data Management and Warehousing Domain Architecture provides high quality, consistent data for Online Transaction Processing (OLTP) and Online Analytical Processing (OLAP) which includes Executive Information Systems (EIS) and Decision Support Systems (DSS).

Proper data management and warehouse architecture will enable the State of Connecticut to leverage the most value from our data assets.

Introduction/Background

There is a distinction between the concepts of "Data" and "Information".

"Data" is stored in multiple applications systems on multiple platforms using multiple methods across the state and is used to perform day-to-day operations. If this distributed data is grouped together in a meaningful format, it can provide valuable "Information" to the state's business organizations, decision-makers and the public.

"Data" is captured using online transaction processing (OLTP) systems to perform mission-critical daily operations. Typically, many users simultaneously add, modify, delete and view data using OLTP applications. OLTP systems are characteristically designed to perform transactions one record at a time.

"Information" is derived from online analytical processing (OLAP) systems used for analysis, planning, and management reporting through access to a variety of sources. An OLAP system usually references information that is stored in a data warehouse. Use of this technology provides the facility to present a comprehensive view of the State enterprise.

"Data" and "Information" are extremely valuable assets to the State of Connecticut. Data Architecture defines an infrastructure for providing high quality, consistent data to be used as the basis for decision support and executive information services as well as traditional transaction applications statewide.

Currently, data is distributed and defined differently by the majority of application systems across the state. Data is often application specific using a variety of formats and semantics. No consistent agency or statewide standards were in place when the applications systems and corresponding databases were developed.

Application systems have historically been monolithic, developed independently from each other. Application development was driven by statutory, policy, or business needs. These applications operate independently from each other. Many do not share any logic or data across system or organizational boundaries. The majority of the data was designed for access by single application systems within a single agency, not for access by multiple application systems in multiple agencies simultaneously.

By promoting the concept of federated data, the state will benefit in the areas of reuse, accuracy, security and currency thus making the data more shareable than the historical monolithic model. The establishment of a metadata repository is an essential method to achieve and maintain federated data.

The Data Management – Data Warehouse Architecture documents the approach for the State to manage its Information and Data. The goals of this architecture are to:

- Separate transaction-processing systems from huge ad hoc queries that are required by analytical, executive decision systems.
- Insulate transaction data systems from the performance and security risk of public Internet inquiries.
- Provide a cross-organizational view of data.
- Promote cross-organizational sharing of data.
- Provide access to data not found in transaction systems such as summary data and historical data.
- Facilitate enhanced end user access and provide more timely answers to business questions by end users.
- Define and disseminate information on the stewardship of data to ensure accuracy, security, privacy and ownership.
- Allow the ability to optimize for performance through separation of OLTP and OLAP.
- Define data consistently across the state using federated data guidelines.

Data Management

Components

Data Architecture defines all the components, interfaces and processes for implementing and managing an integrated, cohesive data policy. These components are defined below:

Data

Data Types

Data types define the domain of values that a data field can have. New technologies are extending the range of data types that can be stored and processed by computers. These offer new ways of interacting and communicating with users and amplify the human/machine interface.

Text and Numeric Fields

Data fields comprising rows of information containing discrete values related to some business entity. Current operational databases are almost completely text and numeric data fields. Since there are discrete values, these can be individually retrieved, queried and manipulated to support some activity, reporting need or analysis. These data types will continue to play a significant role in all our databases.

Images

Scanned pictures of documents, photos and other multi-dimensional forms can be stored in databases. The scanned image is a single data field and is retrieved and updated as a single fact. Software outside of the DBMS is used to manipulate the image.

Geographic Data

Geographic data is information about features on the surface and subsurface of the earth, including their location, shape, description and condition. Geographic information includes spatial and descriptive tabular information in tabular and raster (image) formats. A geographic information system (GIS) is a hardware and software environment that captures, stores, analyzes, queries, and displays geographic information. Typically geographic information is the basis for location-based decision making, land-use planning, emergency response, and mapping purposes.

Multimedia: Voice, Animation and Video

Multi-media applications are increasing as we employ new modalities of communicating with users. Voice can be stored in a database to capture instructional, informative messages that can then be played back rather than displayed as text. This facilitates those situations where keyboards and visual displays are difficult to utilize.

Graphics, animation and video, likewise, offer an alternative way to inform users where simple text does not communicate easily the complexity or the relationships between informational components. An example might be graphic displays of vessels and equipment allowing drill down too more detailed information related to the part or component. Video may be useful in demonstrating some complex operation as part of a training program.

Objects

Objects are composites of other data types and other objects. Objects form a hierarchy of information unlike the relational model. Objects contain facts about themselves and exhibit certain behaviors implemented as procedural code. They also "inherit" the facts and behaviors of their parent objects up through the hierarchy. Relational databases store everything in rows and columns. Although they may support large binary object (LOB) fields that can hold anything, an object database can support any type of data combined with the processing to display it.

Databases

Databases organize data and information into physical structures, which are then accessed and updated through the services of a database management system.

Database (DBMS)

A database is an organization method that links files together as required. In non-relational systems (hierarchical, network), records in one file contain embedded pointers to the locations of records in another, such as customers to orders and vendors to purchases. These are fixed links set up ahead of time to speed up daily processing.

Relational Database Management System (RDBMS)

A relational database management system (RDBMS) is software designed to manage a collection of data. Data is organized into related sets of tables, rows and columns so that relationships between and among data can be established. For example, a vehicle database can contain two tables, one for customer information and one for vehicle information. An "owns" relationship is then established between the two tables.

Multi-Dimensional Databases (MDDBMS)

A multi-dimensional database (MDDBMS) is specifically designed for efficient storage and retrieval of large volumes of data. Multi-dimensional databases are organized into fact tables and dimensions that intersect with the facts table to identify to what the fact pertains. Databases of this construction are used for on-line analytical processing, also known as OLAP.

Data Warehouse – Data Marts

A data warehouse is a database designed to support decision-making in an organization or enterprise. It is refreshed, or batch updated, and can contain massive amounts of data. When the database is organized for one department or function, it is often called a "data mart" rather than a data warehouse.

The data in a data warehouse is typically historical and static in nature. Data marts also contain numerous summary levels. It is structured to support a variety of elaborate analytical queries on large amounts of data that can require extensive searching.

Operational Data Stores

The Operation Data Store (ODS) is a database that consolidates data from multiple source systems and provides a near real-time, integrated view of volatile, current data. An ODS differs from a warehouse in that the ODS's contents are updated in the course of business, whereas a data warehouse contains static data.

Data Access

Data access middleware is the layer of communication between a data access tier and the database. The following components are essential for the data access middleware layer for accessing a relational database in an N-tier application environment:

Structured Query Language (SQL)

A query language used to query and retrieve data from relational databases. The industry standard for SQL is ANSI Standard SQL. RDBMS Vendors implement SQL drivers to enable access to their proprietary databases. Vendors may add extensions to the SQL language for their proprietary databases.

Open database connectivity (ODBC) drivers

Middleware is used to connect database access tools to relational databases using a generic application program interface (API). ODBC drivers are vendor-provided and allow databases to be connected and used by a generic interface. The ODBC drivers enable access to data and provide insulation between a program and the specific RDBMS language used by each database. Database access tools and programs do not have to be customized for each database, because an ODBC configuration file maintains the database connections.

ODBC can be implemented as a client-based solution or a server-based solution.

Processing Access

Access to data falls into two major categories.

Online Analytical Processing (OLAP)

Decision support software that allows the user to quickly analyze information that has been summarized into multidimensional views. Traditional OLAP products, also known as multidimensional OLAP, or MOLAP, summarize transactions into multidimensional views ahead of time. User queries on these types of databases are extremely fast because the consolidation has already been done. OLAP places the data into a cube structure that can be rotated by the user, which is particularly suited for financial summaries.

Online Transaction Processing (OLTP)

Online transaction processing means that master files are updated as soon as transactions are entered at terminals or received over communications lines. It also implies that confirmations are returned to the sender. They are considered ”real-time" systems.

Replication

Replication is used to keep distributed databases up to date with a central source database. Replication uses a database that has been identified as a central source and reproduces the data to distributed target databases. As more and more data is being made available to the public over the Internet, replication of select data to locations outside the firewall is becoming more common.

Replicated data should be accessed by applications in a read-only mode. If updates were allowed on replicated data, data would quickly become corrupted and out of sync. Updates should be directed to the database access tier in charge of updating the authoritative source, rather than to a replicated database.

Replication services are available from most relational database vendors for their particular products.

Replication Services

Replication is the process of distributing information across a network of computers. Replication strategies may also employ some forms of transformation such that the information has different content and meaning. When information has a low volatility, replication may be a valid strategy for optimizing performance.
Replication will need to be evaluated against our network capacity, our volume of activity and local access requirements.

Partial and Full Refresh

A full refresh simply replaces the existing target with a new copy of the source database. It is simple to implement, but may not be practical for large databases due to the amount of time involved in the process of dumping and reloading the data.
A partial refresh replicates only the changes made from the source database to the remote databases. The processing involved in replicating only changes is more complex than a full refresh, but is an optimal solution for a large database. In a partial refresh method, either data or transactions can drive the replication (e.g., sending the exact data that was changed, or sending the same transaction that made the data change on the central database).

Mirroring

Mirroring provides two images of the same database and allows the two databases to be synchronized simultaneously. That is, an update to one causes the "mirror" to also be updated. This form of replication is the most accurate, but also, potentially, the most difficult to achieve and the most costly to operate.

Resource Management

Resource management provides the operational facilities for managing and securing an enterprise-wide, distributed data architecture. It provides a common view of the data including definitions, stewardship, distribution and currency and allows those charged with ensuring operational integrity and availability the tools necessary to do so. Research needs to be done for all components in this category.

Security

Security becomes an increasingly important aspect as access to data and information expands and takes on new forms such as Web pages and dynamic content. Our security policy needs to be examined to ensure that it provides for the new types of databases, the new data types and that it can be enforced given the move to distributed data and internet access.

Administration

Administration encompasses the creation, maintenance, support, backup and recovery and archival processes required in managing a database. We will need the ability to centrally manage all of the enterprise databases to ensure consistency and availability. Distributing data to appropriate platforms will place more importance on administration and control. This becomes the key to maintaining the overall data architecture. Currently database administration is done using the tools and services native to and provided by most relational database vendors for their particular products. Investing in centrally managed administration products and resources will benefit the State in improved data quality, availability and reliability.

Federated Data

Federated data is data that is defined consistently across the state; it is re-useable, shareable, accurate, up-to-date, secure, and managed from a statewide perspective.

Stewardship

The person or group charged with the responsibility of the definition, accuracy, consistency, security and privacy at the elemental level of the data.

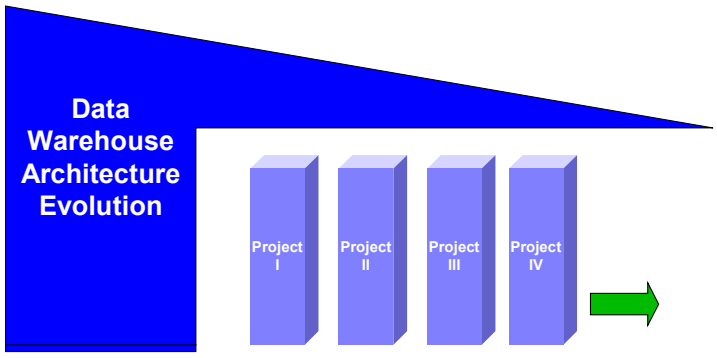
Data Warehouse

What is Data Warehouse Architecture?

Data Warehouse Architecture is a description of the components and services of the warehouse, how they fit together and how they will grow. These descriptions should contain enough information to allow a skilled professional to implement the architecture.

Architecture provides the mechanism to achieve enterprise integration to support State business. It provides an organizing framework that will improve data sharing between agencies, and in the long run allow for faster development, reuse and consistent data between warehouse projects.

Most importantly, this architecture is an evolutionary process. The architecture as defined here was initially developed as a place to start. The first enterprise warehouse projects will be based on this architecture. Increments of additional agency projects will cause this architecture to evolve. As technology changes and improves, that too will most likely require us to make adjustments to this architecture. This incremental development of both the architecture and the warehouse offers an opportunity to learn and to minimize the impact of mistakes.



What is a Data Warehouse?

A data warehouse is something you do, not something you buy. A successful data warehouse does not have an end. Regardless of the methodology, warehousing environments must be built incrementally through projects that are managed under the umbrella of a data-warehousing program. That program will be sponsored and supported at the State of Connecticut Department of Information Technology.

Most of the benefits of the data warehouse will not be realized in the first delivery. The first project will be the foundation for the next, which will in turn form the foundation for the next. Data warehousing at the enterprise level is a long-term strategy, not a short-term fix. It's cost and value should be evaluated across a time span sufficient to provide us with a realistic picture of its cost-to-value ratio.

What a Data Warehouse is NOT.

It is NOT a Data warehouse in a box:

> Many vendors and consultants respond to agency data warehousing needs by offering "data warehousing in a box", or a data mart as their complete end-to-end solution. These pre-packaged data mart solutions sound like an easy way to bypass all the tough issues surrounding the design and integration of a data warehouse. But from the enterprise perspective, the primary value of the data warehouse comes directly from the integration of data. A standalone data mart in each department or agency may independently meet some organizational information needs, but it will not begin to resolve problems that result from creating proprietary islands of information.
>
> From the enterprise perspective, a data mart is not a solution unto itself, it's a component of the overall data warehousing architecture.

It is NOT an Operational Data Store (ODS):

> Operational Data Stores are usually driven by the business need to do faster or more flexible reporting on their operational (transaction) data. There is nothing wrong with building an ODS that will hold a copy of operational

data. Actually, it's quite a common phenomenon. The ODS consolidates data from multiple source systems and provides a near real-time, integrated view of volatile, current data. But an ODS differs from a data warehouse in that its contents are updated in the course of business, whereas a data warehouse contains static data.

ODSs are, however, considered to be a good source of input into a data warehouse because someone has already gone through the hard work of identifying and extracting data from the various legacy systems.

What is Business Intelligence?

It's just the new industry term to refer to information provided by data warehousing.

The Politics of Data Warehousing

The efforts of the Enterprise Warehousing Program will more than likely be slowed by some completely non-technical factors, in particular, the invisible but always present political lines and the politics of data itself. Because data warehouses are infrastructures for sharing, politics and the exercise of power are inherent in all data warehousing projects. We have to except that challenge.

Some data needs to be lawfully protected and the politics in defining and integrating this type of information is understandable. In other cases however, the politics are ungrounded in the thoughts of loss of control over access to the raw data itself and exposing dirty, ambiguous and unflattering data outside an agency. It's understandable that agencies are leery about exposing bad or ambivalent data to inspection by others. It should also be noted that dirty data is present in nearly every commercial and public organization operating today. For fears of this nature, one should note that the reason for the transformation and staging tiers of the warehouse architecture are to clean the data before it's actually loaded into warehouse. We should all accept and get comfortable with the fact that dirty data is a natural occurrence. Making agency data available through the data warehouse gives agencies the opportunity to improve data quality for presenting their data.

Data Sharing

Is it a good candidate for data sharing

Review the criteria below. If the information in the datamart falls into any of these categories, it is a candidate for the Enterprise data warehouse.

- Service to the public will be enhanced or facilitated by sharing the data with others.
- Data is required for another entity to carry out its mission.
- Data is required to support legislative decisions.
- The data already is being shared through another method.
- Someone else will use the data.

Does it meet restrictions and considerations

- Does sharing of this data meet both State and Federal restrictions.
- Are there privacy/confidentiality requirements that need to be considered in the decision.
- Are there usage agreements that must be executed to share this data.
- Data must come from an authoritative source.
- Is the data easily accessible and usable as a data source.

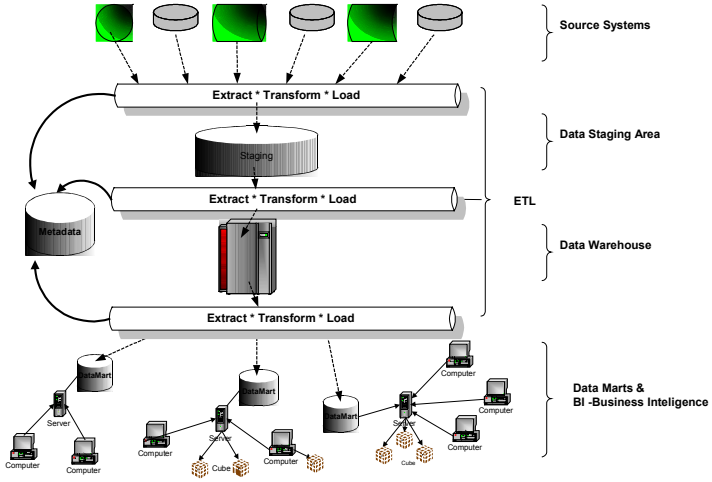
Is there a benefit to sharing this data

- Sharing the data minimizes collection of redundant data that will need cleansing and storage.
- Donors become subscribers.
- Volume of data can be reduced to a manageable level through aggregation.
- Reduces the dependence in IT for reporting.

- Allows the state to leverage scarce resources.
- Provides consistent data between business intelligence projects.
- Allows access to cleansed data.
- Simplifies operational processes of consolidating data from disparate systems.
- Avoids re-keying data from one system to another.

Components

The follow is a list of the seven major components that make up the Enterprise Data Warehouse architecture. These components offer a high level of flexibility and scalability for both the Enterprise and the agencies wishing to implement a Business Intelligence solution.



Source Systems

A data source system is the operational or legacy system of record whose function it is to capture the transactions of the business. There are hundreds of such systems throughout the State of Connecticut. The source systems should be thought of as outside the data warehouse, since we have no control over the content and format of the data. The data in these systems can be in many formats from flat files to hierarchical, and relational RDBMS such as MS Access, Oracle, Sybase, UDB, and IMS to name a few. Other sources of data may already be cleansed and integrated and available from operational data stores. Still other sources are PeopleSoft, Web logs and even external information sources.

Data Staging Area

The data staging area is the portion of the data warehouse restricted to extracting, cleaning, matching and loading data from multiple legacy systems. The data staging area is the back room and is explicitly off limits to the end users. The data staging area does not support query or presentation services. A data-cleansing tool may be used to process data in the staging area to resolve name and address misspellings and the like, as well as resolve other data cleansing issues by use of fuzzy logic.

Data Warehouse Database

The warehouse is no special technology in itself. The data warehouse database is a relational data structure that is optimized for distribution. It collects and stores integrated sets of historical, non-volatile data from multiple operational systems and feeds them to one or more data marts. It becomes the one source of the truth for all shared data.

Data Marts

The easiest way to conceptually view a data mart is that a mart needs to be an extension of the data warehouse. Data is integrated as it enters the data warehouse from multiple legacy sources. Data marts then derive their data from the central data warehouse source. The theory is that no matter how many data marts are created, all the data is drawn from the one and only one version of the truth, which is the data contained in the warehouse.

Distribution of the data from the warehouse to the mart provides the opportunity to build new summaries to fit a particular departments need. The data marts contain subject specific information supporting the requirements of the end users in individual business units. Data marts can provide rapid response to end-user requests if most queries are directed to pre-computed, aggregated data stored in the data mart.

These data marts can reside at DOIT, on local agency servers, or under certain patterns, even on a desktop.

Extract Transform Load

Data Extraction-Transformation-Load (ETL) tools are used to extract data from data sources, cleanse the data, perform data transformations, and load the target data warehouse and then again to load the data marts. The ETL tool is also used to generate and maintain a central metadata repository and support data warehouse

administration. The more robust ETL tools integrate with OLAP tools, data modeling tools and data cleansing tools at the metadata level.

Business Intelligence (BI)

Out of the need to empower the state's user community now emerges the field of Business Intelligence. This space is a key area within the Business Intelligence continuum that provides the tools required by users to specify queries, create arbitrary reports, and to analyze their own data using drill-down and On-Line Analytical Processing (OLAP) functions. Putting this functionality in the hands of the agency's power users allows them to ask their own questions and gives them quick and easy access to the information they need.

One tool however does not fit all. The BI tools arena still requires that we match the right tools to the right end user. The State of Connecticut recognizes that there is a wide spectrum of users within our agencies whose unique needs must be met.

Metadata and the Metadata Repository

A repository is itself a database containing a complete glossary for all components, databases, fields, objects, owners, access, platforms and users within the enterprise. The repository offers a way to understand what information is available, where it comes from, where it is stored, the transformation performed on the data, its currency and other important facts about the data.

It describes the data structures and the business rules at a level above a data dictionary. Metadata has however taken on a more visible role among day-to-day knowledge workers. Today it serves as the main catalog, or map to a data warehouse.

The central metadata repository is an essential part of a data warehouse. Metadata can be generated and maintained by an ETL tool as part of the specification of the extraction, transformation and load process. The repository can also capture the operational statistics on the operation of the ETL process.

Ideally, access to data definitions and business rules in the metadata repository should be end user accessible.

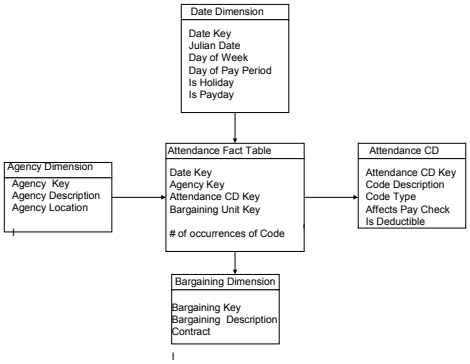
Data Storage Structures

ROLAP:

Relational On-Line Analytical Processing (ROLAP) tools extract analytical data from traditional relational databases structures. Using complex SQL statements against relational tables, ROLAP is able to create multidimensional views on the fly. ROLAP tends to be used on data that has a large number of attributes, where it cannot be easily placed into a cube structure.

MOLAP

Multidimensional On-Line Analytical Processing (MOLAP) is specially designed for the purpose of user understandability and high performance. A multi-dimensional database uses a dimensional model instead of a relational model. A dimensional model is a star schema characterized by a central 'fact' table. One fact table is surrounded by a series of 'dimension' tables. Data is joined from the dimension points to the center, providing a so-called "star". The fact table contains all the pointers to its descriptive dimension tables plus a set of measurements of facts about this combination of dimensions.



HOLAP

Hybrid On-Line Analytical Processing (HOLAP) tools use the best features of multidimensional and relational databases. Relational databases are best known for their flexibility. Until recently relational databases were weak in their ability to perform the same kind of multidimensional analysis that the multidimensional databases are specifically optimized for. The introduction of hybrid relational systems with enhanced abilities to

manipulate star schemas has increased the OLAP capabilities to the relational world. Hybrid tools provide high performance for both general-purpose end users and power users.

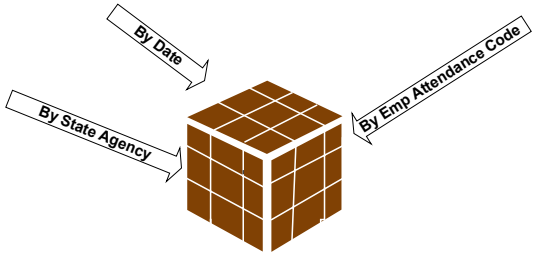
WOLAP

Web-enabled OLAP.

CUBE

In a multidimensional database, a dimensional model is a cube. It holds data more like a 3-D spreadsheet rather than a traditional relational database. A cube allows different views of the data to be quickly displayed. The ability to quickly switch between one slice of data and another allows users to analyze their information in smaller meaningful chunks, at the speed of thought. Use of cubes allows the user to look at data in several dimensions; for example, attendance by Agency, attendance by attendance codes and attendance by date, etc.

Use of a cube can be a far better solution than reviewing a giant report that can be confusing and contains unnecessary additional information or is formatted in a manner that requires additional manual thought or reorganization.



Extraction, Transformation & Load

Transforming data is generally performed as part of the preparation before data is loaded into the data warehouse and data marts. Understanding the business usage of this information and the specific business questions to be analyzed and answered are the keys to determining the transformations necessary to produce the target datamart.

ETL tools are used to extract data from operational and external source systems, transform the data, and load the transformed data in a data warehouse. The same tool is used to extract and transform the data from the warehouse and distribute it to the data marts. When a schedule is defined for refreshing the data, the data extraction and transformation schedule must be carefully implemented so that it both meets the needs of the data warehouse and does not adversely impact the source systems that store the original data.

Extraction

Extraction is a means of replicating data through a process of selection from one or more source databases. Extraction may or may not employ some form of transformation. Data extraction can be accomplished through custom-developed programs. But the preferred method uses vendor-supported data extraction and transformation tools that can be customized to address particular extraction and transformation needs as well as use an enterprise metadata repository which will document the business rules used to determine what data was extracted from the source systems.

Transformation

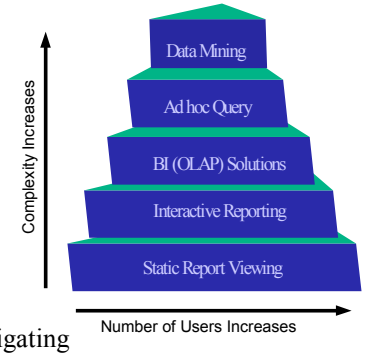
Data is transformed from transaction level data into information through several techniques: filtering, summarizing, merging, transposing, converting and deriving new values through mathematical and logical formulas. These all operate on one or more discrete data fields to produce a target result having more meaning from a decision support perspective than the source data. This process requires understanding the business focus, the information needs and the currently available sources. Issues of data standards, domains and business terms arise when integrating across operational databases.

Data Cleansing

Cleansing data is based on the principle of populating the data warehouse with quality data -- that is, data that is consistent, is of a known, recognized value and conforms with the business definition as expressed by the user. The cleansing operation is focused on determining those values which violate these rules, and, either reject or, through a transformation process, bring the data into conformance.

Data cleansing standardizes data according to specifically defined rules, eliminates redundancy to increase data query accuracy, reduces the cost associated with inaccurate, incomplete and redundant data, and reduces the risk of invalid decisions made against incorrect data.

Business Intelligence Tools



There's a wide choice of tools for business intelligence. There is no single end-user tool that will meet the needs of all categories of users. Selection of multiple types of tools may be needed to support the range of users looking at the information in data marts.

Data Mining

Data Mining tools are a class of products that apply artificial intelligence techniques to the analysis of data. They are about making predictions, not navigating through the data. The promise of data mining tools is that given access to the raw data, the tool can dip through the data looking for patterns and discovering relationships that the user might never have suspected.

Ad hoc Query

Ad hoc query provides business analyst the ability to pose specific questions to produce a result. The ad hoc nature of these queries implies a short shelf life where some situation is being researched or a new opportunity is being explored. The tools falling into this category offer the ability, often through a point and click interface, to search the database and produce a result that can then be displayed, further refined and analyzed. The results can be exported to another desktop application such as Excel, Power point or Word.

OLAP

OLAP tools view information in the form of cubes, or multiple dimensions and allow the user to drill down to lower levels of detail, slice across different dimensions such as time or commodity. These tools are generally used by the business analyst in conducting their research to answer business questions as part of the decision making process.

Interactive Reporting

In interactive reporting, the result set is filtered based on user entered parameter values.

This breed of tool uses a point-and-click display to produce reports with dynamic content.

Static Reporting

Static reporting is a repeatable, pre-calculated and non-interactive request for information. Static reporting is a means of documenting results of such requests in a standard format on a routine basis to a targeted audience. Where reporting of this nature is often viewed as hardcopy, it may take on newer form as the Intranet can become a vehicle for fast dissemination of information.

Principles

These domain principles are intended to guide in the evaluation, selection, design, construction and implementation of this domain and its elements.

Principle 1: Information Is an Enterprise (STATE) Asset

Information is valued as a State asset, which must be capable of being shared.

Rationale

- Enhance and accelerate business Decision-making which requires information beyond the traditional borders of a system or agency.
- Facilitates new enterprise-wide or multi-agency solutions.
- Enhances the efficiency and effectiveness of the delivery of services.
- Expands knowledge of information existence so that its value can be recognized and utilized.
- Increases data's integrity and relevance.
- Enables expanded public access.
- Provides a more global capability to provide information to federal, state, local and general populations.
- Reduces the cost and complexity of business solutions.

Implications

- Need to develop policy pertaining to information stewardship. Requires determination of responsibility for accuracy, access authorization, historical trails, manipulation approval, definitions and integrity relationships.
- Information and its value must be identified by its current keepers. It must be authenticated and documented. Stewardship must be identified or assigned. The meta data must be capable of being universally available so the data contained within can be leveraged by all authorized to use it.
- A mechanism is required to maintain identified meta data information which can be listed, categorized, show stewardship, level of privacy/security and location of information. Need for unified meta data information management to make it accessible for all agencies.
- Need to establish supporting policies for security, privacy, confidentiality and information sharing. Requires Data Use Agreements.
- Data needs to be structured for easy access and management by adopting enterprise data standards.
- Existing identified data should be used from existing sources and not recaptured by new development.
- National standards should be adopted to provide more global sharing capabilities.
- Management tools will be required to maintain and manage a meta data repository.
- Change control procedures need to be defined and adopted to ensure meta data repository is current.
- Methodology is needed to publish and disseminate information on data available for sharing.
- Requires creation of an Enterprise data model.
- Need to establish policy and procedures for maintaining timely and accurate enterprise-wide geographic information.
- Policies for service level agreements need to be defined to determine availability and level of service.
- Need to provide enterprise wide systems that support the creation, storage and retrieval of documents, images and other information rich objects that are used within agency processes or are exchanged with external organizations and constituents.

Principle 2: Architecture Management

The planning and management of the State's enterprise-wide technical architecture must be unified and have a planned evolution that is governed across the enterprise.

Rationale

- Without a unified approach, there will be multiple, and possibly conflicting, architectures.
- Good change requires collaboration and collective planning.
- Architecture must be well thought out.
- Governance will be simplified.
- More stable skill sets will be established.
- Take advantage of volume discounts and economies of scale.

Implications

- A unified approach will require a change in cultural attributes.
- Normal evolution will require prioritization and reprioritization across all IT initiatives.
- Dependencies must be maintained.
- The architecture must be continually re-examined and refreshed.
- Short-term results vs. long term impact must be constantly considered.
- Establishing enterprise architecture takes time and involves a lot of change.
- Make sure that the chosen architecture has a broad range of capabilities to handle vast needs and best of breed solutions in the marketplace (Internet and PDA's and kiosk).
- Planning for retirement of obsolete and non-standard products.
- Requires retraining of staff moving from obsolete technologies.
- Need to develop in-house software engineers, data architects, DBA and Warehouse experts.

Principle 3: Architecture Compliance

Architecture support and review structures shall be used to ensure that the integrity of the architecture is maintained as systems and infrastructure are acquired, developed and enhanced.

Rationale

- To realize the benefits of a standards-based enterprise architecture, all information technology investments must ensure compliance with the established IT architecture.
- For maximum impact, review should begin as early in the solution planning process as possible
- Take advantage of skills and existing architecture.

Implications

- A structured project level review process and authority will be needed to ensure that information systems comply with the IT Architecture and related standards.
- Processes incorporating the principles of this (technical) architecture must be developed for all application procurement, development, design, and management activities.
- This compliance process must allow for the introduction of new technology and standards.
- Conceptual Architecture and Technical Domain principles should be used as evaluation criteria for purchasing as well as developing software.
- Negotiate at the enterprise level to handle increase in compliant systems.
- Need for open mindedness when reviewing for compliance, possible need to broaden existing architecture or considerations for possible exceptions.
- Develop phase out plans
- Inventory of whom is using what architectures.

Principle 4: Leverage Enterprise Data Warehouses

We should leverage a data warehouse and data marts to facilitate the sharing of existing information. This data warehouse will contain the one single version of 'the truth'.

Rationale

- Accelerate and improve decision making at all levels.
- Data can be drawn and combined from multiple agencies without changing the originating systems or developing new systems.
- Reduced business cycle times have led to a need for faster access to more information.
- Reduce the burden on programmers to generate reports and data queries.
- This warehouse will fulfill the need for consistent data.
- Provides cleansed universal data for secured data marts designed to supply information needed by the Internet for public access.
- Reduce the demand on mission critical operational systems.
- Create a user-friendly, centralized location for information.
- Provide the public with more direct access to government information.
- Use as a means to retain data from retired systems.
- Use as a gateway to enterprise-wide geographic information.

Implications

- Data warehousing must become a core competency of IT.
- Data warehousing both requires and supplies configuration standards that need to be developed and maintained.
- End-user tools must be provided to relieve the burden on programmers to provide this functionality.
- End users become more knowledgeable about the information available to them.
- End users become more aware of and knowledgeable of the tools they need to access and analyze it.
- The processes and procedures refreshing the data warehouse will require high levels of reliability and integrity.
- Warehousing is not meant to replace transaction applications shortcomings.
- Guidelines on maintaining data and data retention need to be developed.
- Not all requests for data are simple in nature and appropriate for end user tools.
- Not all data will be available to all users.
- End users should be able to access the data without knowledge of where it resides or how it is stored.
- Data warehouse architecture design requires an integrated design effort to provide usefulness. Full potential of a date warehouse will not be realized unless there is full participation throughout the enterprise.
- User community must be made aware of the (un-) timeliness of information.
- There is still a governance requirement for the process of data retrieval to prevent monopolization.
- Data warehouses serve as Web-based gateways to commonly used enterprise-wide public and geographic information.

Principle 5: Ensure Security, Confidentiality and Privacy

IT systems should be implemented in adherence with all security, confidentiality and privacy policies and applicable statutes.

Rationale

- Helps to safeguard confidential and proprietary information.
- Enhances public trust.
- Enhances the proper stewardship over public information.
- Helps to ensure the integrity of the information.

Implications

- Need to identify, publish and keep the applicable policies current.

- Need to secure data elements.
- Categorize access control, this may vary depending on the data and the audience.
- Need to monitor compliance to policies.
- Must make the requirements for security, confidentiality and privacy clear to everyone.
- Education on issues of privacy and confidentiality must become a routine part of normal business processes.
- Audit capabilities of data access.
- Understanding that part of the stewardship role interprets security, confidentiality and privacy.
- All access requests for data that is not publicly available should be made to the steward of the data.
- Requires a means to publish and implement changes to the status of data's access requirements.

Principle 6: Reduce Integration Complexity

The enterprise architecture must reduce integration complexity to the greatest extent possible.

Rationale

- Increases the ability of the enterprise to adapt and change.
- Reduces product and support costs.
- Take advantage of existing architecture patterns based on proven designs.

Implications

- Decreases the number of vendors, products, and configurations in the State's environment.
- Must maintain configuration discipline.
- Will sacrifice performance and functionality in some instances.
- Will rely on components supplied by vendors, which will make the enterprise more vulnerable.
- Need to factor cost of vendor dependency when figuring the total cost of ownership.
- Determination of "the greatest extent possible" includes consideration of how reducing complexity can negatively impact providing critical client services.

Principle 7: Re-use before Buying, Buy before Building

We will consider re-use of existing tools and infrastructure before investing in new solutions

Rationale

- Use and availability of effective packaged solutions is increasing.
- Using tested solutions reduces risks.
- Reduces the total cost of ownership.
- Allows use of familiar tools or technology at the agency business intelligence level.

Implications

- Software license agreements and system development contracts should be written to allow for re-use across State government.
- "The definition of "reusable" will include solutions available from other government entities (e.g., other states, federal government, etc.).
- Areas that provide clear advantages and businesses cost savings are likely to require quick adaptation.

Principle 8: Integration

Systems must be designed, acquired, developed, or enhanced such that data and processes can be shared and integrated across the enterprise and with our partners.

Rationale

- Increase efficiency while better serving our customers (e.g., the public, agencies, etc.).
- Redundant systems cause higher support costs.
- Ensures more accurate information, with a more familiar look and feel.
- Integration leads to better decision making and accountability.
- Make better use of existing data.

Implications

- IT staff will need to consider the impacts on an enterprise wide scale when designing applications.
- Will need a method for identifying data and processes that need integration, when integration should take place, whom should have access to the data, and cost justification for integration.
- It will be necessary to coordinate, maintain and arbitrate a common set of domain tables, data definitions, and processes across the organization.
- Over integration can lead to difficult data management and inefficient processes.
- Use of meta data repository
- Enterprise integration teams comprised for dedicated enterprise data architects and applications architects are required to assist in integration efforts.
- Stewardship review of integration.
- There is a need to evaluate the practicality of an integrated project before development.

Principle 9: Reengineer First

New information systems will be implemented after business processes have been analyzed, simplified or otherwise redesigned as appropriate.

Rationale

- There is no real "value" in applying technology to old, inefficient processes.
- Work processes will be more streamlined efficient and cost effective.
- Work processes, activities, and associated business rules will be well understood and documented.
- Reduces the total cost of ownership.

Implications

- Need to have an agreed upon business re-engineering process.
- Need to identify the business need for data.
- Need to determine the legal requirement for retention of data.
- New technology will be applied in conjunction with business process review.
- Business processes must be optimized to align with business drivers.
- Additional time and resources will have to be invested in analysis early in the systems life cycle.
- Organizational change will be required to implement reengineered work processes.
- May require regulatory or legislative change.

Principle 10: Total Cost of Ownership

Adopt a total cost of ownership model for applications and technologies which balances the costs of development, support, training, disaster recovery and retirement against the costs of flexibility, scalability, ease of use, and reduction of integration complexity.

Rationale

- Consideration of all costs associated with an Architecture over its entire life span will result in significantly more cost effective system choices.
- Enables improved planning and budget decision-making.

- Reduces the IT skills required for support of obsolete systems or old standards.
- Simplifies the IT environment.
- Leads to higher quality solutions.

Implications

- The State budget process needs to accommodate Total Cost of Ownership of an Architecture over a longer timeframe than current budgeting models.
- Will require looking closely at technical and user training costs especially when making platform or major software upgrades during the lifetime of the system.
- Requires designers and developers to take a systemic view.
- Need to selectively sub-optimize individual IT components.
- Need to develop a cost of ownership model.
- Need to ensure coordinated retirements of systems.
- Need to consider budget issues for staffing and training.
- Need to develop a cost structure for providing access to shared information.
- Need to provide funding for data costs that are not billable or recoverable.
- Need to establish permanent, reliable funding mechanisms for developing enterprise-wide geographic information such as aerial photography, satellite imagery, transportation, and hydrographic data layers.

Principle 11: Shared Components Using an N-tier Model

Infrastructure and data access will employ reusable components across the enterprise, using an n-tier model.

Rationale

- You can make significant changes to a component of a system, such as changing from a Windows client to a web-browser client, without changing the rest of the system.
- Enables simplification of the environment and geographical independence of servers.
- Takes advantage of modular off-the-shelf components
- Reuse will lower costs and maintenance efforts.
- Allows for leveraging skills across the enterprise.

Implications

- Component management must become a core competency.
- Requires development of a culture of reuse.
- Design reviews become crucial.
- Data marts can be modularized without making components too small or too simple to do useful “work”.

Principle 12: Logical Partitioning and Boundaries

The logical design of application systems and databases should be highly partitioned. These partitions must have logical boundaries established, and the logical boundaries must not be violated.

Rationale

- A change in a database or application can potentially affect many large programs, if they are not highly partitioned.
- You can not separate the components in a system from each other without creating logical boundaries.
- Re-coding leads to time-consuming re-testing.

- Partitioning isolates/minimizes change impact.
- Partitioned code is more adaptive to changes in internal logic, platforms, and structures.

Implications

- Applications need to be divided into coded entities (e.g., presentation, process, and data access).
- For databases, there will be a need to develop competency in partitioning horizontally and vertically; this will result in more but simpler tables and views
- Design reviews must ensure logical boundaries are kept intact.
- Increases data management responsibilities for DBAs.
- Requires increased analytical skills of project analyst to determine when partitioning takes place based on expected data and or access requirements.

Principle 13: Physical Partitioning of Processing

We should separate on-line transaction processing (OLTP) from data warehouse and other end-user computing and Internet access.

Rationale

- Separating end-user requests and OLTP maximizes the efficiency of both environments.
- Growth in OLTP is incremental, and requirements are predictable.
- Growth in data warehouses and end-user computing has been nonlinear, and requirements are very difficult to predict.
- Internet access ramifications are increasingly difficult to predict and can be erratic.
- Fosters the concept of data stewardship.
- Take advantage of designs best suited for the type of access required.
- Easier to secure data inside and outside the firewalls.

Implications

- Data marts represent a type of configuration standard for physical partitioning.
- Data warehousing and data marts must become core competencies of IT.
- Business and IT must agree on the purpose and objective of the data warehouses.
- Data redundancy will be necessary.
- Data marts will not reflect the most current data.
- It is not always necessary or even desirable to physically partition data such as when there is a low scalability requirement.

Principle 14: Mainstream Technologies

IT solutions will use industry-proven, mainstream technologies.

Rationale

- Avoids dependence on weak vendors.
- Reduces risk.
- Ensures robust product support.
- Enables greater use of commercial-off-the-shelf solutions.
- A more knowledgeable workforce.
- More availability of training and consulting services.

Implications

- Need to establish criteria for vendor selection and performance measurement.
- Need to establish criteria to identify the weak vendors and poor technology solutions.

- Requires migration away from existing weak products in the technology portfolio.
- Will reduce some solution choices.
- Need to respond as changes in technology occur.

Principle 15: Industry Standards

Priority will be given to products adhering to industry standards and open architecture.

Rationale

- Avoids dependence on weak vendors.
- Reduces risks.
- Ensures robust product support.
- Enables greater use of Commercial-off-the-Shelf solutions.
- Allows flexibility and adaptability in product replacement.
- System migration will be easier.

Implications

- Requires a culture shift.
- Need to establish criteria to identify standards and the products using them.
- IT organizations will need to determine how they will transition to this mode.
- Will reduce some solution choices.

Principle 16: Disaster Recovery / Business Continuity

An assessment of business recovery requirements is mandatory when acquiring, developing, enhancing or outsourcing systems. Based on that assessment, appropriate disaster recovery and business continuity planning, design and testing will take place.

Rationale

- Due to factors such as the Internet and Y2K, customers and partners have heightened awareness of systems availability.
- The pressure to maintain availability will increase in importance. Any significant visible loss of system stability could negatively impact our image.
- Continuation of business activities without IT is becoming harder.
- Application systems and data are valuable State assets that must be protected.

Implications

- Systems will need to be categorized according to business recovery needs (e.g. business critical, non-critical, not required).
- Alternate computing capabilities need to be in place.
- Systems should be designed with fault tolerance and recovery in mind.
- Plans for work site recovery will need to be in place.
- Costs may be higher.
- Data must be capable of on-line backups to provide 24 x 7 availability.

Principle 17: Scalability

The underlying technology infrastructure and applications must be scalable in size, capacity, and functionality to meet changing business and technical requirements.

Rationale

- Reduces Total Cost of Ownership by reducing the amount of application and platform changes needed to respond to increasing or decreasing demand on the system.

- Encourages reuse.
- Leverages the continuing decline in hardware costs.

Implications

- Scalability must be reviewed for both “upward” and “downward” capability.
- May increase initial costs of development and deployment.
- Will reduce some solution choices.

Best Practices

DATA

It is important to ensure that data is backed-up, recoverable and available. Currently, most database backups are done using the tools and services native to and provided by most relational database vendors for their particular products.

Although backups are covered in the Platform Architecture chapter, these are recommended best practices pertaining to database backup and availability strategies.

Best Practice 1:

The greater the percentage of changes to the data, the more frequent your backups should be.

Rationale

- There is a high cost associated with having to replace and re-enter data.

Best Practice 2:

Consider the availability requirements when determining the backup strategies.

Rationale

- If access to the data requires 24 X 7 availability then online backup strategies are required.
- With online backups, the database can be backed up while users or applications are connected. Database performance can be decreased and the backup may take longer.
- With offline backups, no users or applications can be connected to the database however the backup may be faster.
- Always consider the down time associated with restoring older backups and longer recovery logs vs. more frequent backups with smaller recovery logs.

Best Practice 3:

Always use transaction logging on databases that require roll-forward recovery to point of last transaction.

Rationale

- In this method, changes made to the database are retained in logs. You first restore the database using a backup image; then you use the logs to reapply changes that were made to the database since the backup image was created.

Best Practice 4:

When using transaction logging on databases, put the logs on a different disk.

Rationale

- In the event of a disk failure on the disk containing a database, the database can be restored from the last backup and the log is still available for roll forward recovery.

Best Practice 5:

For data requiring 24 x 7 availability, consider using fail over, mirroring, or other high availability strategies.

Rationale

- Allows for uninterrupted data access.

DATABASES

Best Practice 1:
Segregate OLTP data and OLAP data into separate databases.

Rationale
- OLTP databases are used by online users for mission critical day-to-day operations. Segregating OLTP data and OLAP data in separate databases reduces the impact of ad hoc and large queries from decision support systems.
- Separating OLTP and OLAP data also aligns the Data Architecture with the Application Architecture, thus separating decision support applications from operation support applications.

Best Practice 2:
Centralize data that needs to be shared and current.

Rationale
- High-volume transaction data shared across locations and needing to be current for all locations should be centralized so all locations have access to the live data.
- Replicating frequent updates to distributed databases increases systems complexity and network traffic.
- Data should be centralized when one or more of the following criteria occur:
    - Many users need access to latest changes (i.e., OLTP systems).
    - There is a lack of skills and tools at multiple sites to manage distributed data.
    - There is a need to provide a consolidated and integrated database for federated data on an open platform.

Best Practice 3:
Design databases to be modular, not monolithic.

Rationale
- This practice provides better performance for backup and recovery, better transaction performance due to parallelism (e.g., a complex request can be broken down and be processed by multiple databases at the same time), higher reliability, availability, and scalability.

Best Practice 4:
Design for all replicated data to be read-only.

Rationale
- Updates should occur through the source where the data originates to facilitate the ease of data management.
- The exception is when it is clearly indicated that a discrepancy in data updates at distributed locations would not adversely affect the business.
- If a distributed site has a requirement to make updates, direct all updates back to the authoritative source.

Best Practice 5:
Conform to State of Connecticut Standards for databases.

Rationale

- The Department of Information Technology (DOIT) for the State of Connecticut, under the authority granted to the Chief Information Officer in Sec. 4d2. of the Connecticut General Statutes, is establishing database technology standards to:
  - Promote portability and interoperability of database application programs.
  - Facilitate maintenance of database systems among heterogeneous data processing environments.
- Follow EWTA patterns for DB2, UDB, Oracle, and SQL Server.

Best Practice 6:

New databases implemented should use a relational DBMS supporting ANSI Standard SQL (currently SQL92).

Rationale

- The State of Connecticut standard for the syntax and semantics of SQL language and for accessing SQL based relational databases is ANSI S3.135-1992 Database Language SQL as delimited by FIPS PUB 127-3.

- The SQL implementation and relational database products must support database security using the following database access controls: GRANT and REVOKE privilege facilities, the VIEW definition capabilities, and some Discretionary Access Control (DAC) mechanisms.
- Nonstandard language features shall be used only when the needed operation or function cannot reasonably be implemented with the standard features alone.
- Relational databases offer dependability, flexibility, and compatibility for future data needs.
- Data can be maintained and readily accessed through standard SQL calls. SQL is an industry standard for the data access tier of an application system and for data access tools.

DATA WAREHOUSE

It is important to design a data warehouse to answer the business and performance requirements of the business problem. These are the recommended best practices pertaining to designing and implementing the data warehouse.

Best Practice 1:

Begin data mart efforts by addressing a specific requirement for a specific decision support application, keeping growth and scalability in mind.

Rationale

- This practice is similar to data warehouse design, but the tools used should be designed to support the business intelligence requirements.

Best Practice 2:

Identify specific requirements for data availability, freshness (i.e., live, 24 hours old, etc.), and recoverability.

Rationale

- Some data in the data warehouse needs to be refreshed more frequently than others. If the original data is fairly stable and not as volatile, the data warehouse may only need daily, weekly, or even monthly loads.
- When the original data is frequently changing or is more volatile, it may be necessary to consider an ODS as an alternative.

Best Practice 3:

Ensure appropriate security is applied to both data marts and the data warehouse.

Rationale

- Prevent unauthorized users from tampering with the data.
- Certain data stored in the data warehouse may be sensitive. Providing protection for privacy and confidentiality of data must be considered.
- Data should only be put from the data warehouse to the data marts under the control of the data warehouse.
- Subscribers cannot transfer rights and data to another subscriber.
- Must define subscription and data use agreements.
- Agencies are responsible for non-sharable data never reaching the data warehouse.

Best Practice 4:
Choose a data warehousing project manager who is business process oriented rather than technology oriented.

Rationale

- A business process oriented manager ensures that the data warehouse will meet the business needs of the end users.
- The data warehousing project manager must manage the expectations and sponsorship of the data warehouse.
- The data warehousing manager can make sure the data is easy to use and understand.

Best Practice 5:
Assess the source data that will populate a data warehouse for accuracy, quality, and veracity.

Rationale

- Data needs to be accurate to ensure good business decisions.
- Data needs to be relevant to the business need and consistent across multiple sources.
- Data must be complete. It must contain the information necessary to answer the data warehouse business need.
- The data assessment also involves evaluating the business rules associated with that data. The appropriate business rules must be applied to the data to maintain accuracy.

Best Practice 6:
Allow read-only access to end users of the data warehouse.

Rationale

- Updates should only occur through the OLTP source where the data originates.

Best Practice 7:
Direct all information queries against data marts, not OLTP databases. Conversely, operational transactions should be directed to operational databases only, not OLAP databases.

Rationale

- Data marts contain data that has been checked for consistency and integrity, and represent a cross-functional view of data.
- OLTP transactions should not interact with a data warehouse or data marts

Best Practice 8:
Establish the enterprise data warehouse as the authoritative source for all data marts.

Rationale

- Data administration policies, procedures and tools are required.
- Project design reviews are required. A mandatory deliverable must be identification of all data sources for the project.

Best Practice 9:
Perform periodic validity audits against the data warehouse information model to ensure a high level of confidence in the quality and integrity of the data.

Rationale

- Accelerated decision making requires high quality data. If operational data has changed or additional data is needed, changes must be made in the information model and in the data warehouse itself.
- The data stored in a data warehouse should conform to the information model.
- The source data populating a data warehouse should be verified for consistency and accuracy.
- The data warehouse should correspond to business needs.
- Ensuring the integrity and quality of data is the responsibility of both the business users and IS.

Best Practice 10:
The implementation plan should match critical business needs.

Rationale

- Start with strategic business needs.
- Optimize critical data access before less-critical data access.
- Optimize high-volume data access before low-volume data access.
- Optimize applications and data with high-service-level requirements before those with lower requirements.

Best Practice 11:
Consider the network when designing OLAP systems.

Rationale

- The network is a partner to data marts and can impact performance and how the data mart is designed.
- Estimate the impact on the network when designing data marts.

Best Practice 12:
Ensure data entry quality is built into new and existing application systems to reduce the risk of inaccurate or misleading data in OLTP systems and to reduce the need for data hygiene.

Rationale

- The system should be designed to reject invalid data elements and to assist the end user in correcting the entry.
- All updates to an authoritative source OLTP database should occur using the business rules that own the data, not by direct access to the database.

Best Practice 13:
During data warehouse design, determine the logic needed to convert the data, plan and generate the extraction and transformation routines, and quality assure the data populating the data warehouse.

Rationale

- Planning for data extraction and transformation should start at the same time the data warehouse design starts.

- Data extraction and transformation is an important process for populating the data in a data warehouse and for ensuring that the data in a data warehouse is accurate.
- The data warehouse will contain data from disparate operational data sources. This data will need to be cleaned in the staging area to resolve any differences before it can be stored for analysis in the data warehouse.
- Data extraction and transformation logic includes data conversion requirements and the flow of data from the source operational database to the data warehouse, and subsequently to the data mart.

Best Practice 14:
Develop a schedule for data extraction that both meets the needs of the data warehouse users and does not impact an OLTP system.

Rationale
- Evaluate the impact of data extraction to any OLTP systems accessed.
- Business process requirements should determine frequency of extracts.

Best Practice 15:
Document data extraction and transformation information in the metadata repository.

Rationale
- Data extraction and transformation are important aspects of a data warehouse. This provides the information map connecting the data populating a data warehouse with its source operational databases.

DATA REPLICATION
It is important to address data replication to answer the business and performance requirements of the business problem.

Best Practice 1:

Replication of data should be based on needs such as availability, security, performance, or decision support.

Rationale
- Data quality and integrity is more manageable when replicated and distributed data are read-only.
- Replication can insure uninterrupted access to critical data.
- Replication can isolate production data from external users.
- Replication can facilitate load balancing through synchronization of distributed databases.
- Replication allows separation of OLTP data from information required for decision support without degrading performance of source systems.

Best Practice 2:
Replicated data should be read only whenever possible.

Rationale
- Read only access eliminates the need for multidirectional replication.
- One way replication is easier to implement and requires less planning.
- One way replication requires less system resources.

METADATA REPOSITORY

It is important to address the metadata repository to maximize the long-term value of data warehousing and to facilitate data sharing.

Best Practice 1:

Actively maintain a metadata repository for data warehousing.

Rationale

- The repository contains metadata, or information about the data, in the data warehouse.
- The repository represents the shared understanding of the organization's data.
- The repository can be built incrementally, in stages, based on data warehouse design and implementation.
- The repository should support multiple types of data elements, such as graphics.
- Changes in the repository must occur before, and correspond to, the changes to the data warehouse environment.

Best Practice 2:

The repository management system used to store metadata should be built with the same ETL tool that is used to integrate data input to the enterprise warehouse.

Rationale

- Metadata references many common data elements used by multiple application systems.
- If the repository databases use the same ETL tool, metadata will easily transfer between repositories to build the information model.

Standards

The State of Connecticut's current database inventory has been documented in the following matrices; each technology item was categorized as follows:

Technology Usage Categories

Obsolete Standards

It is highly likely that these standards or products, while still in use, will not be supported by the vendor (industry, manufacturer, etc.) in the future. Some products and standards have already reached the non-supported state. Plans should be developed by the agencies or the State to rapidly phase out and replace them with strategic standards or products. No development should be undertaken using these standards or products by either the agencies or the State.

Transitional Standards

These are standards or products in which an agency or the State has a substantial investment or deployment. These standards and products are currently supported by DOIT, the agencies, or the vendor (industry, manufacturer, etc.). However, agencies should undertake development using these standards or products only if there are no suitable alternatives that are categorized as strategic. Plans should be developed by the agencies or the State to move from transitional to strategic standards or products as soon as practical. In addition, the State should not use these standards or products for development.

Note: many older versions of *strategic* standards or products fall into this category, even if not specifically listed in a domain architecture document.

Strategic Standards

These are the standards and products selected by the state for development or acquisition, and for replacement of *obsolete* or *transitional* standards or products. (Strategic means a three to four year planning horizon.) When more than one similar strategic standard or product is specified for a technology category, there may be a preference for use in statewide or multi-agency development. These preferred standards and products are indicated where appropriate.

Note: some strategic products may be in "pilot testing" evaluation to determine implementation issues and guidelines. Pilot testing must be successfully completed prior to full deployment by the agencies or the State.

Research / Emerging Standards

This category represents proposed strategic standards and products that are in advanced stages of development and that should be evaluated by the State. The sum of these standards or products may already be undergoing "hands-on" evaluation. Others will need to be tracked and evaluated over the next 6 to 18 months.

Product Standards

Database – Mainframe

Table 1 Mainframe Database Products

Product or Tech Standard	Obsolete	Transitional	Strategic	Research
IBM DB2			X	
Dms-2200		X		
Focus		X		
IMS		X		

Product or Tech Standard	Obsolete	Transitional	Strategic	Research
Progress		X		
RMS		X		
SAS			X	
Unisys MAPPER 2200		X		
Unisys RDMS 2200		X		
VSAM		X		
Wang/Pace	X			
VAX ORACLE RDB		X		

Database – Server

Table 2 Server Database Products

Product or Tech Standard	Obsolete	Transitional	Strategic	Research
FOCUS		X		
FOXPRO		X		
MS ACCESS		X		
ORACLE			X	
SAS			X	
SQL SERVER			X	
SYBASE		X		
TURBOIMAGE		X		
IBM DB2 (UDB)			X	
Unisys MAPPER NT		X		
Unisys COOLICE		X		
DBASE		X		
CLIPPER		X		
ESRI ARCSDE			X	
SHAPEFILE / COVERAGE		X		
ESRI GEO DB			X	

Database – DESKTOP

Table 3 Desktop (PC) Database Products

Product or Tech Standard	Obsolete	Transitional	Strategic	Research
DATAEASE	X			
ENABLE	X			
FILEMAKER PRO		X		

Product or Tech Standard	Obsolete	Transitional	Strategic	Research
FIRSTCHOICE	X			
FoxPro		X		
LOTUS APPROACH		X		
MS ACCESS			X	
Oracle 8				X
PARADOX		X		
SMART	X			
SAS			X	
IBM DB2 (UDB)				X

Data Modeling

Table 4 Data Modeling Products

Product or Tech Standard	Obsolete	Transitional	Strategic	Research
ERWIN				X
Oracle Designer				X
Powerdesigner				X
UML				X

Database – Connectivity

Table 5 Database Connectivity Standards

Product or Tech Standard	Obsolete	Transitional	Strategic	Research
ADO			X	
DAO		X		
DB2 CONNECT-DRDA			X	
JDBC			X	
ODBC			X	
OLE DB			X	
RDO		X		
SYBASE NATIVE		X		
SQL NET			X	
Oracle NET8			X	
NEON				

Backup – Recovery

Table 6 Backup and Recovery

Product or Tech Standard	Obsolete	Transitional	Strategic	Research
DB2 UTILITIES			X	
ORACLE UTILITIES			X	
IMPORT/ EXPORT			X	
SQL SERVER UTILITIES			X	
SYBASE BACKUP		X		
VAX RMU UTILITY		X		
ARCSERVE DB			X	

ETL

Table 7 Extract Transform & Load Products

Product or Tech Standard	Obsolete	Transitional	Strategic	Research
INFORMATICA				X

To Be Determined

To meet the mission of the Data Management and Data Warehouse Domain Architecture, initiatives have been identified to give direction toward accomplishment of the domain goals.

Development of a Federated Metadata Repository

- Provide agencies with the ability to share their metadata (database related information) with other state agencies.
- Enhanced statewide application collaboration and interoperability.
- Provide a centralized source of information about data for the enterprise.
- Allow the state to leverage its data resources.

Initiate a Stewardship Program

- Identify data stewards and data trustees of key shared data.
- Develop solutions to ensure accuracy, privacy and appropriate sharing.
- Identify the roles, responsibilities, and procedures (for both release and notification) for responding to FOI and other data requests.

Provide Data warehousing capabilities

- Allow for replicated and combined data from multiple agencies without changing the originating system.
- Create a user friendly, centralized location for information & decreased burden on programming staff.
- Reduce demand on mission critical operational systems by separating transaction processing from analytical processing.
- Develop data conversion and Data Warehouse retention standards

- Create Federated data by creating uniform data elements and values; cleanse and convert data from transaction-based databases into Federated Data Warehouse.
- Include ability to allocate costs of Data Warehouses maintained by central (rather than agency) organizations.

Data Architects

- Cultivate a skill set for architects with data expertise in specific areas of business
- Provide a resource with an enterprise picture of our valued data assets.

Research and select tool sets

- Data modeling
- Warehouse development
- Data Cleansing & Transformation Tools